Universitat de Lleida

# TREBALL FINAL DE GRAU


Estudiant: **Humbert Vallés Teixidó**

Titulació: **Grau en Enginyeria Informàtica**

Títol de Treball Final de Grau: **Manteniment predictiu d'una bomba de calor mitjançant analítica de dades**

Director/a: **Jordi Planes Cid i Josep Pijuan Parra**

Presentació
Mes: Setembre

Any: 2018

M’agradaria donar les gràcies a tota la gent que m’ha ajudat d’una forma o altra durant aquest projecte. Primer de tot, als meus dos tutors: en Jordi Planes per part de la UdL i en Josep Pijuan per part d’Eurecat. Gràcies per la vostra ajuda, suport i paciència guiant-me durant el desenvolupament d’aquest projecte.

També m’agradaria donar les gràcies a la resta de l’equip d’Eurecat Lleida, per oferir-me la possibilitat de treballar amb ells en un ambient en el qual he pogut créixer com a professional i com a persona.

A la Clàudia, per ajudar-me en els moments d’estres, sempre disposada a fer-me veure que no n’hi ha per tant i que ho sóc capaç de fer-ho.

I finalment, a la meva família, pel seu suport incondicional durant aquests anys. En especial, a la meva mare, la qual no ha deixat mai de confiar en mi i a qui segurament dec una disculpa per tot l’estres que li he provocat durant aquests darrers mesos. Sense vosaltres, res d’això hauria estat possible.

# Abstract

L'objectiu d'aquest projecte és analitzar un conjunt de dades pertanyents a una bomba de calor i aplicar tècniques de manteniment predictiu per detectar anomalies en el funcionament que precedeixin una fallada d'algun element del sistema.

L'autor ha seguit un model de recerca utilitzat en el camp de l'analítica de dades per tal d'intentar obtenir models que permetin predir el comportament de la bomba de calor i així poder detectar futures fallades i actuar en conseqüència per tal d'evitar-les.

Aquest document mostra el procés seguit durant la recerca que s'ha dut a terme i els resultats obtinguts a partir de la recerca, a més d'una introducció al camp del manteniment predictiu i de la metodologia de desenvolupament d'un projecte d'analítica de dades.

# Índex

# Índex d'imatges

# Capítol 1

# Introducció

Segons prediccions de Dell EMC, a l'any 2020 la quantitat de dades que generarem i copiarem arribarà als 44 zettabytes anuals. Es a dir, 44 trilions de gigabytes [1].

L'augment en l'ús de dispositius IoT, d'emmagatzematge en el núvol i l'ús d'aplicacions que cada vegada ofereixen més funcionalitats que requereixen d'una quantitat de dades elevada per funcionar correctament, fa que resulti bàsic desenvolupar i millorar aquelles tècniques que permetin classificar, analitzar i donar ús a aquestes dades de forma eficient i obtenint resultats que tinguin un impacte real en la millora del camp en el qual s'apliquin.

La ciència de les dades o *data science* és un dels camps que ha nascut per fer front a aquests reptes. La seva finalitat és la d'agrupar tècniques extretes de camps com l'estadística, les matemàtiques i la programació, combinar-les amb una capacitat de pensar de manera enginyosa, de mirar les coses de forma diferent i de ser capaç de netejar, preparar i tractar dades i obtenir així un camp que ens permetrà poder contestar aquelles preguntes que ens ofereixin les dades [1].

No obstant, en els darrers temps s'està convertit en el que en anglès s'anomena una *buzzword*: una paraula que es fa molt popular i que deriva d'un terme tècnic però que passa a ser utilitzada amb un significat diferent o més laxe que l'original amb la intenció d'impressionar als altres.

Aquest creixent interès general pel camp de les data science ha arribat a tal punt en que un article publicat en la Harvard Business Review la considerava la feina més sexy del segle

[1] Dell EMC Global Data Protection Index II: https://www.emc.com/microsites/emc-global-data-protection-index/index.html

XXI [2].

Així doncs, intentant allunyar-nos el més lluny possible d'aquesta laxetat en l'ús del terme data science, aquest treball buscarà servir dos propòsits: el primer, és el d'actuar com una introducció en el camp de les data science, específicament en la branca de l'analítica de dades mitjançant un cas d'ús pràctic que permeti posar en funcionament els coneixements teòrics que s'obtindran durant la realització d'aquest treball per tal de crear una experiència el més propera possible a un projecte real.

El segon, és generar una base de coneixement i unes pautes que puguin ser utilitzades a posteriori per encarar el desenvolupament de projectes similars en els quals s'analitzin series temporals adreçat a la detecció d'anomalies en sistemes tèrmics. Com per exemple el projecte LowUP [3].

LowUP, és un projecte europeu on es combinen diferents sistemes tèrmics amb tecnologia innovadora per una gestió energètica més eficient. L'objectiu de LowUP, és desenvolupar i demostrar el funcionament de tres noves tecnologies altament eficients: un sistema d'escalfament, un sistema de fred i un sistema de recuperació de calor.

Un dels socis d'aquest projecte és Eurecat, el Centre Tecnològic de Catalunya. L'objectiu d'Eurecat en el projecte LowUP és similar al que tenim en aquest projecte: desenvolupar un model que permeti detectar anomalies en les diferents bombes de calor i altres màquines que formaran part del sistema.

Obtenint aquesta base de coneixement prèvia, s'espera poder facilitar les futures tasques d'anàlisi i tractament de les dades que s'hagin de realitzar un cop els diferents equips del projecte LowUP es trobin en funcionament.

Per tal de complir aquests propòsits, el problema a tractar serà l'estudi d'un conjunt de dades o *data set* públic que conté lectures de diversos sensors situats en una bomba de calor. Les lectures tenen lloc cada minut i cobreixen un període de tres anys: del 2011 al 2013.

Així doncs, es tracta d'un data set d'una mida considerable però sense arribar a un volum que requereixi d'eines específiques per tractar amb grans quantitats de dades. D'aquesta

[2]Data Scientist: The Sexiest Job of the 21st Century: https://hbr.org/2012/10/data-scientist-the-sexiest-job-of-the-21st-century

[3]LowUP Project: http://lowup-h2020.eu/about-lowup/

forma, disposem de dades suficients per poder treballar sense preocupació per una possible falta de dades, però sense els problemes d'eficiència i requeriments a nivell tan de hardware com de software derivats del treball amb quantitats de dades molt grans.

## 1.1 Objectius

Per tal d'assolir els propòsit esmentats en la introducció, es buscarà complir els objectius següents:

1. Obtenir una base de coneixement general dels conceptes relacionats amb les Data Science
2. Obtenir una base de coneixement profunda de les tècniques i metodologies utilitzades en l'analítica de dades
3. Dur a terme un procés d'estudi de les dades i ser capaç d'identificar quines dades són rellevants i quines són errònies, defectuoses o irrellevants.
4. Identificar les diferents parts de la bomba i determinar a quina part de la bomba correspon cadascun dels sensors dels quals disposem dades
5. Crear un algorisme que correlacioni el consum energètic i la potència utilitzada per la bomba amb la temperatura ambient utilitzant el data set del qual disposem
6. Detectar senyals que precedeixin una fallada de la bomba de calor
7. Monitorar i identificar possibles disminucions en l'eficiència de la bomba de calor
8. Crear un algorisme genèric que donats els paràmetres de funcionament de la bomba sigui capaç d'identificar possibles fallades.

## 1.2 Estructura del projecte

Aquest document es troba estructurat en 5 capítols.

El *Capítol 1* correspon a la introducció d'aquest projecte.

En el *Capítol 2* es revisa l'estat de l'art del manteniment predictiu i de les metodologies per realitzar analítica de dades.

El *Capítol 3* presenta les tecnologies que s'han emprat per dur a terme el projecte.

En el *Capítol 4* s'explica en detall el model que s'ha seguit per desenvolupar el projecte,

Finalment, el *Capítol 5* presenta les conclusions del projecte i les futures línies de treball.

# Capítol 2

# Estat de l'art

En aquest capítol, es presentarà una descripció de les tecnologies que s'utilitzen per a tractar problemes de manteniment predictiu i s'introduirà el model CRISP-DM, el qual ha estat seguit durant el desenvolupament del projecte. Aprofito per deixar clar que l'estructuració i una part significativa de les descripcions del manteniment predictiu provenen de [7].

## 2.1 Manteniment predictiu

El manteniment predictiu pot ser dividit en tres tècniques diferents depenent de l'origen de les dades que utilitzem:

- Manteniment basat en els sensors existents: són aquells mètodes en els quals les dades que s'utilitzen són les que s'obtenen de la lectura dels sensors situats en el interior de la màquina. Aquests sensors, són els que mesuren variables com poden ser la temperatura, la pressió, el cabal, el volum, etc. És a dir, utilitzem els sensors ja implantats de base no només per conèixer l'estat de la màquina en tot moment sinó també per verificar paràmetres com les calibracions i el temps de resposta del sensor de cara a detectar anomalies.

- Manteniment basat en els sensors de testeig o de diagnosi: aquesta segona categoria utilitza les dades obtingudes de sensors de testeig com poden ser acceleròmetres per tal de mesurar vibracions i sensors acústics per detectar fugues.

- Manteniment basat en mesuraments actius de les senyals de testeig: mentre que els altres dos mètodes són passius, ja que depenen de sensors que ja es troben situats en la màquina, aquest tercer mètode dependrà de senyals que seran injectades en l'equipament per tal de testejar-lo. Aquest grup inclou mesuraments actius com poden

ser tests d'aïllament elèctric, o un test LCR, en el qual es mesura la inductància, capacitància i resistència elèctrica de l'equipament. Aquest mètodes s'utilitzen per detectar esquerdes, corrosió o desgast de l'equipament.

Encara que la finalitat i les característiques de les dades que recollirem variaran clarament depenent del mètode que utilitzem, no existeix cap algorisme predefinit per treballar en la detecció de les anomalies en cadascun dels mètodes. L'algorisme a utilitzar, dependrà completament de les dades que tinguem.

## 2.2 Cross-industry standard process for data mining

Cross-industry standard process for data mining (CRISP-DM) és model de procés obert que defineix una metodologia general a aplicar a l'hora de realitzar projectes de minat de dades.

Tal i com hem comentant anteriorment, els camps relacionats amb les Data Science són relativament nous, i com a tals encara es troben en un procés constant d'evolució i desenvolupament. Així doncs, tot i que la majoria disposen de certes bones pràctiques acceptades, no existeix un model estandarditzat que determini quina és la millor manera de procedir en el desenvolupament d'un projecte.

No obstant, en el camp de l'analítica de dades CRISP-DM s'està convertint en un estàndard *de facto* de com desenvolupar projectes. En el diagrama següent es mostren les 6 fases que defineixen el funcionament del model:

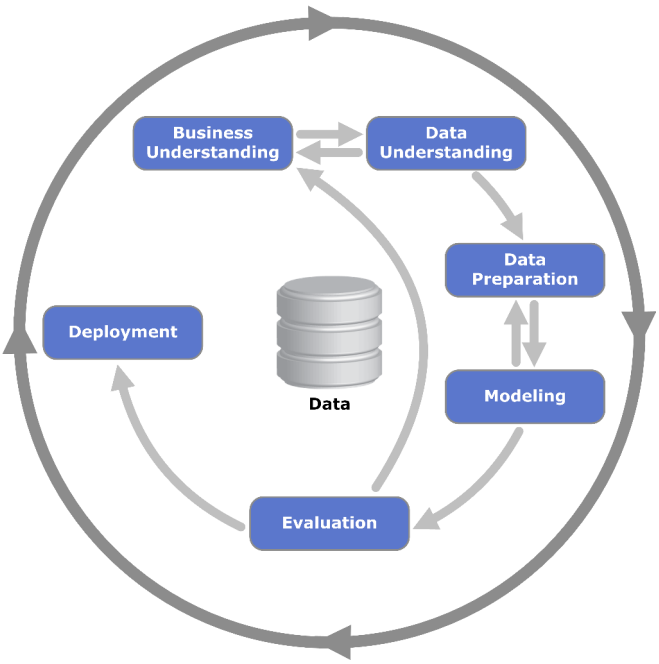


Fig. 2.1 Diagrama de procés de CRISP-DM

Com es pot veure, es tracta d'un model cíclic en el qual es pot tornar endarrere en tot moment, ja que en els projectes d'anàlisis de dades, a mesura que aprofundim en les dades i obtenim majors coneixements, és normal que es revaluïn els objectius i els procediments a aplicar i que siguin modificats o fins i tot substituïts per altres que reflecteixin la nova informació de la qual disposem.

De la mateixa manera, encara que el projecte estigui finalitzat i s'hagi desplegat la solució, caldrà retornar al model cada cert temps per garantir que continua funcionant correctament o per modificar-lo d'acord amb les noves dades que podem haver obtingut. A grans trets, el diagrama anterior es correspon amb als passos següents:

- Business Understanding: Definir el problema que s'intenta solucionar i determinar quins són els objectius finals i com es compliran.
- Data Understanding: Examinar les propietats generals de les dades de les quals disposem (format, quantitat, identificadors dels camps, etc) i determinar si són suficients per poder complir els objectius marcats.
- Data Preparation:
  - Data Cleaning: Analitzar les dades per tal d'identificar punts dèbils i inconsistències.

  - Data exploration: Visualitzar les dades i formular hipòtesis sobre el problema que busquem solucionar.
  - Feature Engineering: Determinar quines són les característiques (aquelles dades que siguin rellevants) més importants del data set, i obtenir característiques addicionals a partir de les dades de les quals disposem.
  - Data Split: Determinar quina és la millor manera de dividir les dades entre dades d'entrenament i dades de testeig de cara a passar-les al model que desenvoluparem.

- Modelling: Crear un model, entrenar-lo, validar-lo i avaluar la seva eficiència
- Deployment: Desplegar el model desenvolupat. En alguns casos, l'anàlisi realitzat i els coneixements obtinguts és l'únic que ens interessa i per tant no serà necessari desplegar cap solució, però sí que resultarà important mostrar-los de forma gràfica i presentar de forma entenedora els resultats que hem obtingut per tal de realitzar la transferència del coneixement obtingut.

# Capítol 3

# Tecnologies emprades

La complexitat d'aquest projecte rau en l'anàlisi i en el tractament de les dades, per la qual cosa la gran majoria del treball a realitzar es centrarà en això, la qual cosa significa que les tecnologies que s'utilitzin han d'actuar com a suport i no requerir una atenció especial per la seva complexitat d'ús o per que no existeix una base prèvia de coneixement sobre el seu funcionament.

Més enllà d'això, les tecnologies que s'han emprat acostumen a ser la base de molts dels projectes d'analítica de dades que es realitzen en entorns professionals, ja que ofereixen el que estem buscant: un suport que ens permeti tractar amb les dades amb agilitat i sense preocupar-nos per la complexitat de l'eina que hem d'utilitzar.

## 3.1 Python

Python ha estat el llenguatge de programació escollit per al desenvolupament del projecte. Python és un llenguatge interpretat d'alt nivell amb una filosofia que fa èmfasi en fer codi que sigui fàcilment llegible.

En el camp de les Data Sciene, existeixen un conjunt de llibreries de les quals parlarem a continuació que aporten totes les eines necessàries per tal de desenvolupar un projecte. I aquestes llibreries són per a Python. És per aquest motiu que s'ha escollit Python com a llenguatge de programació del projecte.

## 3.2 Git i Github

Pel control de versions del projecte s'ha decidit treballar amb Git. L'elecció es deu a que és model de control de versions amb el qual he treballat i per tant amb el que em sento còmode, més enllà del fet de que Git s'ha convertit en un estàndard *de facto* en versions de control.

Pel que fa a Github, és una plataforma web que utilitza Git i que permet compartir codi fàcilment amb altres desenvolupadors a més d'oferir eines que permeten gestionar projectes. De nou, el fet d'haver treballat anteriorment utilitzant Github ha portat a que sigui escollit com a eina per implementar el control de versions.

## 3.3 Anaconda

Anaconda és una distribució de software lliure i codi obert de Python i R per a aplicacions relacionades amb data science i aprenentatge automàtic. La seva finalitat principal és facilitar la gestió de paquets, i ho fa mitjançant el seu propi gestor de paquets anomenat conda. A més, la seva aplicació d'escriptori anomenada Anaconda Navigator proporciona un gestor d'entorns virtuals.

Aquestes característiques la fan ideal pels nostres requisits, ja que ens permet començar a treballar de forma ràpida en qualsevol sistema que ens puguem trobar i ens facilitarà l'accés i l'instal·lació d'altres aplicacions i/o llibreries que s'utilitzaran en el desenvolupament d'aquest projecte.

## 3.4 Jupyter Notebook

Jupyter Notebook és una aplicació web de codi obert que ofereix un entorn de computació interactiu online que permet crear i compartir documents que continguin codi, equacions, gràfiques, text, etc.

Al oferir un entorn interactiu i al qual es pot accedir fàcilment, resulta especialment útil per realitzar tasques que requereixin visualitzar dades per tal de netejar-les o transformar-les, per desenvolupar models estadístics, per crear gràfiques, etc.

En resum, cobreix totes les necessitats que tenim en aquest projecte pel que fa a una plataforma sobre la qual programar. És per això que ha estat seleccionada com l'eina en la qual s'ha escrit tot el codi.

## 3.5 NumPy

NumPy és una llibreria per a Python que conté els elements fonamentals per realitzar activitats científiques de computació. Ofereix suport pel tractament d'arrays i matrius multidimensionals de grans dimensions a més d'una gran quantitat de funcions matemàtiques d'alt nivell. La seva versatilitat i eficiència converteixen aquesta llibreria en un element bàsic de tot projecte d'analítica de dades.

## 3.6 Pandas

Pandas és una altra llibreria per a Python per manipular i analitzar dades. Entre altres, ofereix la possibilitat de treballar amb DataFrames: objectes que permeten manipular dades amb índex integrat. També ofereix eines per llegir i escriure en diferents formats de fitxer, eines per tractar dades que faltin, modificar estructures de dades afegint-hi o eliminant-n'hi columnes, funcions per treballar amb series temporals, etc.

En resum, ofereix totes les funcionalitats necessàries pel nostre projecte, a més de ser una llibreria altament optimitzada i que per tant ens donarà molts millors resultats que no pas si intentem programar nosaltres mateixos un element que pandas ja cobreix.

## 3.7 Scikit-learn

Scikit-learn és una llibreria de software lliure que ofereix eines d'aprenentatge automàtic. Entre altres, ofereix algorismes de classificació, regressió i clustering com poden ser Support vector machines, random forests, gradient boosting, k-means i DBSCAN.

Així doncs, ens ofereix una gran quantitat d'algorismes dels quals se n'han acabat utilitzant diversos durant el desenvolupament del projecte.

## 3.8 Llibreries de visualització de dades

Pel que fa a la visualització de gràfiques, s'han utilitzat tres llibreries diferents: Seaborn, Pyplot i Matplotlib. La primera, ens permet crear gràfiques d'una gran qualitat i altament modificables però requereix cert coneixement i agilitat amb l'eina ja que les comandes per generar gràfiques acostumen a ser més complicades.

Pyplot és una llibreria que permet generar gràfiques interactives de forma senzilla i requerint molt menys codi que altres llibreries com poder ser Seaborn.

Finalment, Matplotlib és la més senzilla de les tres llibreries i és la que porten de base els notebooks amb els quals treballarem.

# Capítol 4

# Desenvolupament del projecte

En aquest capítol es descriurà el procés que s'ha seguit d'anàlisi del data set seguint el model CRISP-DM. Primer, es descriurà el data set i s'explicarà l'estudi previ de les dades que s'ha realitzat. A continuació, s'explicaran les diverses formes en les quals s'ha treballat amb les dades de cara a intentar complir els objectius marcats i finalment es presentarà el model que s'ha dissenyat per tal de detectar anomalies.

## 4.1 Business Understanding

En aquesta primera secció de desenvolupament del model de CRISP-DM, es busca definir el problema que s'intenta solucionar i determinar quins són els objectius del projecte i com s'assoliran.

No obstant, abans d'entrar en la definició del problema i dels objectius, considero que és important mencionar un element clau en el desenvolupament de qualsevol projecte d'analítica de dades i del qual malauradament no hem disposat en aquest projecte: es tracta de la figura de l'expert del domini.

L'expert del domini és una persona amb coneixement amplis i profunds del camp en el qual es realitza l'anàlisi i que actua com a guia de quins camins s'han de seguir i quines són les preguntes i les característiques rellevants que s'haurien de tenir en compte. Així doncs, acostuma a ser una de les persones que ajuden a definir el problema a resoldre i els objectius a complir.

Tot i que en aquesta primera part del projecte considerem que no ha suposat un problema gran no disposar-n'hi, sí que ha afectat en gran mesura el desenvolupament d'altres parts del

projecte i és per això que s'ha considerat necessari mencionar-ho.

Retornant de nou al contingut de la secció en si mateix i tal com ha estat comentat anteriorment, els objectius que es busquen poder assolir amb aquest projecte es poden dividir en dues categories: la primera és el manteniment predictiu i la segona el disseny d'algorismes i/o models genèrics. A continuació es mostren les decisions preses en cadascuna de les categories.

### 4.1.1 Manteniment Predictiu

El nostre objectiu és poder detectar un possible fallada abans de que es produeixi i poder planificar una aturada del sistema per resoldre el problema amb el menor cost possible. Per tal de determinar que un element de la bomba no funciona adequadament es consideraran dues possibilitats:

1. Es detecten anomalies en les lectures dels sensors.
2. Es redueix l'eficiència de la bomba.

Pel que fa a les anomalies, el primer que hem de ser capaços de determinar és què es considera una anomalia. Hem de ser capaços de discernir entre una lectura errònia del sensor i un valor anòmal provocat per un mal funcionament de la bomba. A partir d'aquí, s'han de poder trobar patrons en les anomalies que ens indiquin quan l'error és prou significatiu com per poder decidir que aquella part de la bomba hauria de ser revisada.

I pel que fa a l'eficiència de la bomba de calor, existeixen tres maneres de determinar si s'ha reduït. La primera, consisteix en analitzar el Coefficient of performance (COP) de la bomba [1]. El COP es calcula dividint la calor útil que ha generat la bomba entre el seu consum. El seu valor normal s'acostuma a situar entre 2 i 6. Per exemple, un COP de 4 significaria que per cada kWh d'energia elèctrica consumit, la bomba ens proporciona 4kWh d'energia tèrmica. Per tant, si podem identificar un reducció en el llarg del temps del COP de la bomba, significa que s'ha de dur a terme una tasca de manteniment.

La segona forma de comprovar l'eficiència de la bomba consisteix en analitzar la *Delta T*: la diferència de temperatura entre l'entrada i la sortida del sistema. Igual que amb el COP, la Delta T teòricament, s'hauria de mantenir constant. Per tant, en cas de que es redueixi,

[1] COP: https://en.wikipedia.org/wiki/Coefficient_of_performance

significa que és necessari realitzar una tasca de manteniment en la bomba.

Finalment, la última forma de comprovar l'eficiència consisteix en desenvolupar un model capaç de correlacionar la temperatura ambient amb el consum elèctric i la potència utilitzada per la bomba. Si podem generar un model capaç de predir amb precisió el consum esperat de la bomba, en cas de que el consum real sigui més alt i aquesta diferència augmenti en el temps, significa que l'eficiència de la bomba s'està reduint.

### 4.1.2 Algorismes i models d'ús general

El segon grup d'objectius, és el disseny d'un algorisme o un model genèric o parametritzable que pugui ser utilitzat per altres bombes de calor, com poden ser les pertanyents al projecte LowUP, amb les mínimes modificacions possibles o únicament canviant els paràmetres de l'algorisme.

En aquest projecte, aquesta segona categoria finalment ha estat deixada de banda i no s'ha treballat de cara a desenvolupar-la, ja que el projecte ja ha suposat un repte suficient per ell mateix com per intentar extrapolar un model o un algorisme d'ús genèric.

## 4.2 Data Understanding

En aquesta secció, s'examinaran les propietats generals de les dades de les quals disposem i es determinarà si són suficients per poder encarar el desenvolupament dels objectius marcats.

### 4.2.1 Dades utilitzades

Pel desenvolupament d'aquest projecte s'ha utilitzat un data set provinent d'un dels reptes que es van proposar a la hackaton "Minds and Machines Berlin Hackathon 2017", concretament el "Electrification Challenge". El data set s'anomena "Heatpump Timeseries data set" i fou proporcionat per la European Heat Pump Association (EHPA) i l'empresa Fraunhofer. [6].

### 4.2.2 Característiques del data set

El data set conte les dades corresponents a les lectures dels diversos sensors d'una bomba de calor situada a Alemanya. Es proporcionen les dades d'operació de tres anys, del 2011 al 2013, amb una lectura cada minut. Les dades es troben dividides en tres fitxers *.csv* (un per

cada any). A més, també s'inclou un fitxer en el qual es descriuen les característiques de les dades en anglès i en alemany.

El format del fitxer és el mateix per a tots tres anys. Les dues primeres columnes contenen el dia i l'hora en que es realitza la lectura dels sensors en format ISO 8601 estès [2]. És a dir: AAAA-MM-DD i hh:mm:ss. La resta de columnes (un total de 152) contenen les dades obtingudes pels diferents sensors.

En total, disposem aproximadament d'un milió i mig d'entrades per cadascuna de les columnes. No obstant, quasi la meitat contenen un gran nombre de valors nuls, ja que per la gran majoria de les columnes de dades existeix una segona columna amb el mateix contingut però en la qual s'aplica un filtre que només deixa les dades obtingudes mentre la bomba de calor es troba en actiu i passa la resta de lectures a nul.

El ID de cadascun dels paràmetres dels quals disposem està format per un seguit de sigles que descriuen a quina part del sistema pertany cada paràmetre, quin tipus de dada conte, la unitat de mesura en la que es troba i en alguns casos, un filtre que ens indica si la lectura es fa únicament quan la màquina està encesa o si inclou també les lectures durant els períodes d'inactivitat. Aquestes sigles es troben en alemany.

Després d'aquest anàlisis general les característiques del data set, es determina que es disposen de dades més que suficients per tal de poder encarar el desenvolupament del projecte i poder complir els objectius.

## 4.3 Data Cleaning

L'objectiu d'aquesta part del projecte és analitzar les dades per tal d'identificar punts dèbils i inconsistències i mirar de suplir-les.

No obstant, abans de dur a terme aquest procés es va decidir traduir les sigles que formen els IDs dels paràmetres de l'alemany a l'anglès creant així una nova nomenclatura per tal de facilitar el tractament i la comprensió de les dades amb les quals treballem. En l'apèndix A es mostra una taula amb aquesta nova nomenclatura i què representen cadascuna de les sigles.

[2] ISO 8601: https://www.iso.org/iso-8601-date-and-time-format.html

Un cop modificats els noms, es va iniciar el procés de Data Cleaning. Aquesta part del projecte va ser la primera en la qual la manca d'un expert del domini va ser rellevant. Per tal de determinar quines dades són simplement errors de lectura per part del sensor o el contingut de quines columnes és irrellevant és necessari un coneixement previ del camp en el qual es realitza l'analítica de les dades. En el nostre cas, no hem disposat un expert del domini que ens ho pogués indicar.

A més, com que l'objectiu principal del projecte és fer manteniment predictiu i gran part d'aquest procés és ser capaç de detectar anomalies, si es procedís a sanejar les dades que contenen lectures fora del normal considerant que es tracta d'errors del sensor, es podrien estar eliminant dades que indiquen una fallada del sistema, les quals necessitem, i per tant es va prendre la decisió de no dur a terme cap procés de sanejament de les dades.

No obstant sí que es van eliminar les 10 últimes columnes del conjunt de paràmetres. Aquestes 10 columnes, contaven totes amb el prefix *zzy* en el seu ID i el seu valor oscil·lava entre 0, 1 i 2. Com que en el fitxer de descripció de les dades no s'explicava què significava aquest prefix, es va decidir eliminar-les per evitar treballar amb paràmetres dels quals no podíem determinar la seva funcionalitat.

## 4.4 Data exploration

En aquest apartat és quan realment es comença a treballar amb el contingut de les dades. El procés que es va realitzar consisteix en estudiar el contingut de les dades a traves dels seus paràmetres estadístics i visualitzar-les mitjançant gràfiques. A partir d'aquest estudi, es pretén formular hipòtesis sobre quines dades poden ser més rellevants i per tant ens poden resultar útils de cara a tractar el problema que busquem solucionar.

La primera part dels procés d'exploració de les dades va ser carregar les dades en un Notebook de Jupyter. Es va decidir treballar amb les dades corresponents a l'any 2011, ja que encara no s'havia creat un únic fitxer que contingués les dades dels tres anys i es va considerar que les dades d'un any eren més que suficients per fer aquesta exploració prèvia. Això també va permetre reduir el cost en temps de computació de les operacions que s'haurien de realitzar portant a que el procés fos molt més àgil.

El següent pas fou analitzar les estadístiques principals de cada paràmetre: nombre de dades, màxim, mínim, mitjana i 25,50 i 75 percentils. Això ens va donar una idea general de les distribucions de les dades en els diferents paràmetres. Mentre que alguns eren molt

estables i centrats al voltant de la mitjana, altres mostraven una alta variabilitat. El fet de tractar amb els màxims i els mínims també va servir per identificar alguns paràmetres que tenien valors extrems que es desviaven molt del comportament habitual de les altres dades del seu grup.

A continuació es va procedir a crear un Scatter plot per tal de poder detectar correlacions a simple vista entre els diferents paràmetres del data set. Al tenir tants paràmetres, el procés es va fer per parts. El que segurament va tenir més rellevància va ser el gràfic en el qual es comparaven els diferents paràmetres amb el filtre ON, és a dir, que només tenen dades quan la bomba de calor està activa. Aquesta gràfica es troba en l'apèndix B

Tot i que el gran nombre de gràfiques que es troben en l'Scatter plot pot portar a confusions, si s'observa en profunditat es pot observar que existeix correlació directa entre diferents paràmetres. Per tal de confirmar aquesta correlació, es va procedir a calcular el coeficient de correlació utilitzant una funció de la llibreria Scikit-learn.

A partir d'aquestes correlacions, es va poder observar que diversos paràmetres estaven altament correlacionats amb la temperatura ambient, la qual cosa era d'esperar donat que el funcionament de la bomba canviarà depenent del temps.

També es va observar que diversos paràmetres amb un sufix comú estaven correlacionats, donant-nos una idea per tant de que diferents part del sistema amb un prefix diferent i que per tant haurien de pertànyer a diferents parts de la bomba tenen una font d'entrada comú que els afecta. Per altra banda, també es va poder observar una clara correlació entre diferents paràmetres els quals tenien un mateix prefix en el seu ID.

També es va procedir a representar els diferents paràmetres en gràfiques utilitzant la seva temporalitat en l'eix de les x. D'aquesta manera, es va poder veure clarament que existia un factor d'estacionalitat que afectava el funcionament normal de la bomba. Tot i que intuïtivament ja es podia suposar que existirien diferències en el funcionament de la bomba depenent de l'època de l'any, aquestes gràfiques van servir per confirmar-ho.

A més, també es van identificar factors d'estacionalitat diaris a partir de l'estudi de parts més petites del data set, ja que al tractar amb tot un any de dades els detalls de la gràfica quedaven coberts i resultava impossible identificar aquest tipus de patrons menors.

## 4.5 Feature Engineering

L'objectiu d'aquesta part del projecte és determinar quines són les columnes de dades més importants del data set, i un cop identificades, obtenir característiques addicionals a partir d'elles per tal de poder-les utilitzar a posteriori en la part de Modelling.

En un projecte real, en el qual es disposés d'un expert del domini, normalment es començaria estudiant aquells paràmetres que l'expert considerés rellevants, creant així una línia mestra a seguir de cara a explorar les dades cercant aquelles més rellevants.

En el nostre cas, al no disposar d'un expert del domini, s'ha obert l'oportunitat d'investigar lliurement les dades sense cap tipus de prejudici previ que ens indiqués on mirar. Aquesta major llibertat de moviments ens ha obligat a actuar de forma autònoma, utilitzant la nostra intuïció a cada pas del del camí i a fer un ús extens i profund del coneixement que obteníem de les dades per tal de poder seguir endavant.

És per això que crec que és important mencionar que, tot i que pugui semblar s'han anat provant diferents idees sense cap tipus de rumb, la realitat és que les diferents tècniques que s'han provat han estat escollides de forma lògica i sempre aconsellats. No obstant, la manca de resultats concloents pot portar a donar una impressió de desordre i manca de rigorositat, però aquesta impressió no pot estar més lluny de la realitat.

### 4.5.1 Càlcul del COP

El primer que es va intentar, va ser calcular el COP de la bomba de calor, de cara a poder determinar el COP normal de funcionament de la bomba i així poder buscar patrons que ens indiquessin si hi hi havia una reducció al llarg del temps.

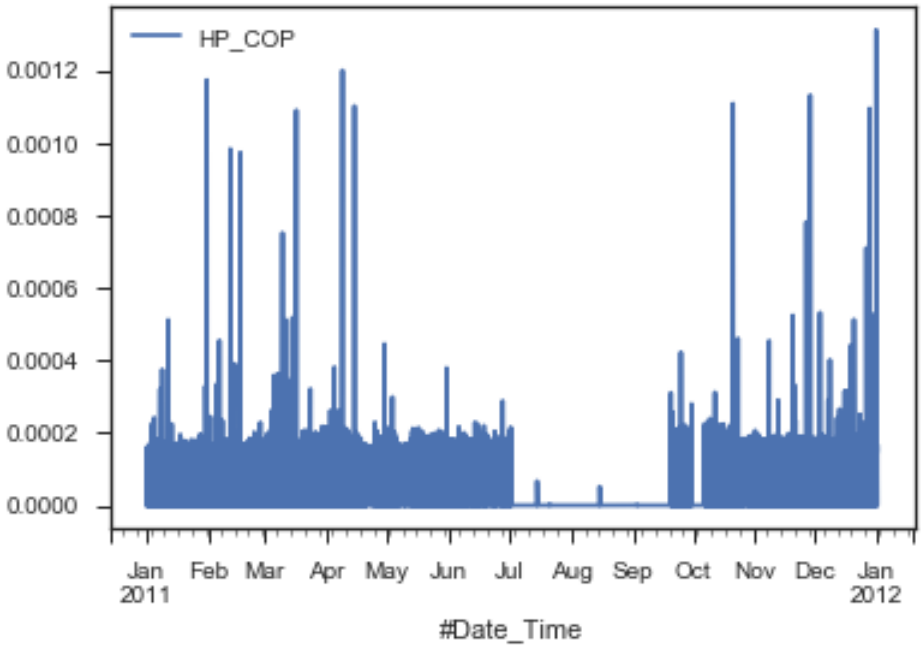


Fig. 4.1 Resultats del càlcul del COP de la bomba de calor

Com es pot observar clarament, ens vam trobar amb dos problemes molt greus: el primer i més important, era que els nostres càlculs estaven 4 ordres de magnitud per sota del que es consideraria un COP normal. El segon, és que encara que el primer problema fos un simple error de conversió d'unitats en el càlcul, el fet que el COP fluctués de forma bastant elevada ens indicava que molt probablement no seria possible detectar una reducció del COP que ens permetés afirmar que l'eficiència del sistema estava baixant.

### 4.5.2 Estudi del consum energètic

A continuació, es va decidir explorar la mateixa via però aquesta vegada buscant un augment en el consum energètic que indiques també una disminució de l'eficiència.

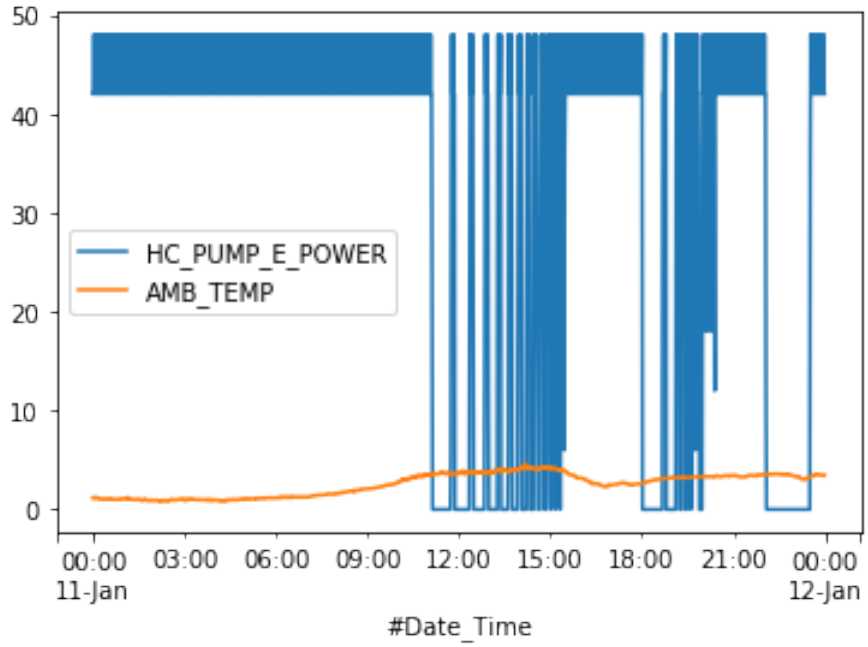


Fig. 4.2 Consum de la bomba de calor en un dia

Primer de tot, explicar que el fet de que la gràfica només mostri el consum d'un únic dia és perquè s'ha considerat més important una major claredat en la gràfica que no pas una representació completa del resultats. No obstant, els resultats mostrats en aquesta gràfica són molt similars als que s'obtindrien si s'agafes tot l'any de dades.

Retornant a la recerca en sí, tal com es veure en la gràfica, el consum energètic sempre es troba entre 40 i 50 i no tendeix a desviar-se'n a no ser que la bomba s'aturi. Aquest comportament es manté en la resta de dades de la columna. És per això que es va decidir deixar de banda l'estudi del consum energètic ja que no semblava mostrar cap tipus de variació en el seu comportament al llarg del temps.

### 4.5.3 Creació d'esdeveniments d'activació de la bomba

Després d'aquestes dues primeres aproximacions fallides, es va decidir canviar de paradigma i passar a tractar només amb aquelles columnes per les quals existís un filtre d'activació de la bomba (tal com s'ha explicat anteriorment, aquelles columnes en les quals només hi ha dades quan la bomba de calor està encesa).

Un cop identificades les columnes que tenien un filtre, es van crear el que van passar a ser anomenats esdeveniments d'activació. Un esdeveniment està format pel conjunt de dades que es llegeixen des de que s'engega la bomba de calor fins que s'apaga.

A continuació, es van generar paràmetres addicionals per tal d'obtenir una major quantitat d'informació sobre els diferents esdeveniments per tal d'intentar determinar si existia algun patró de funcionament que els marqués, però no se'n va trobar cap.

### 4.5.4 Clusterització

Un cop finalitzat aquest procés, es va decidir aplicar un procés de clusterització de cara a identificar grups de paràmetres que treballessin junts per tal de poder determinar quins eren els paràmetres més importants.

Per fer aquesta clusterització, es va decidir utilitzar l'algorisme k-means, ja que és ràpid d'executar i la llibreria Scikit-learn n'ofereix una implementació.

La primera clusterització que vam realitzar es va basar en el nombre d'esdeveniments d'activació que tenien lloc a diari i el primer problema amb el que ens vam trobar fou que certs paràmetres tenien un nombre molt elevat d'activacions.

Això provocava que l'algorisme crees dos clústers: un amb el paràmetre dissonant i l'altre amb la resta d'elements, ja que la diferència entre qualsevol d'ells era molt menor que amb el paràmetre dissonant. A més, al augmentar el nombre de clústers, com que la resta de clústers seguien sent molt propers, la puntuació obtinguda per cada model i que s'utilitzava per determinar el nombre òptim de clústers empitjorava i ens resultava impossible saber quants grups existien realment.

És per això que va ser necessari eliminar els paràmetres que provocaven aquestes clusterització errònies i repetir el procés fins que no apareixien clústers d'un sol membre. Per tal de determinar el nombre òptim de clústers es van utilitzar dos sistemes de valoració: el primer la puntuació que donava l'algorisme k-means en sí i el segon un mètode anomenat Silhouette [3]. Com que ambdós sistemes ofereixen una representació gràfica de la puntuació obtinguda, va resultar relativament senzill comparar-les i determinar el nombre òptim de

[3]Silhouette: https://en.wikipedia.org/wiki/Silhouette_(clustering)

clústers.

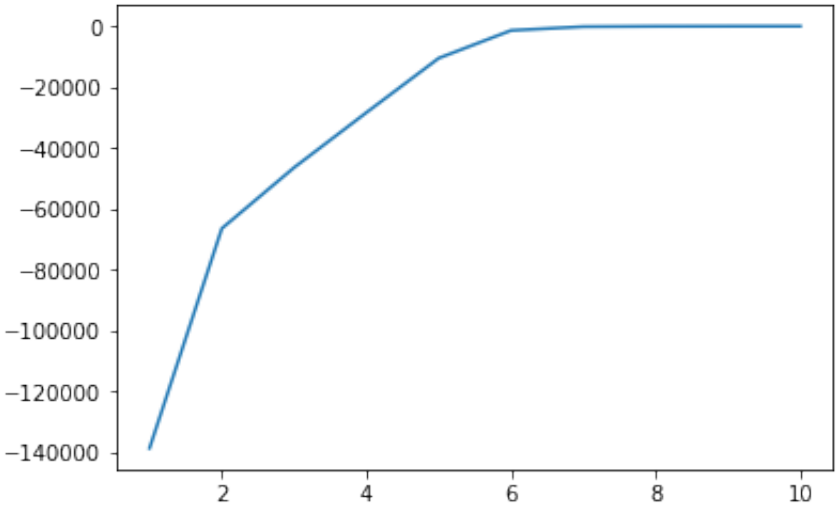


Fig. 4.3 Puntuació d'un model donada per K-means

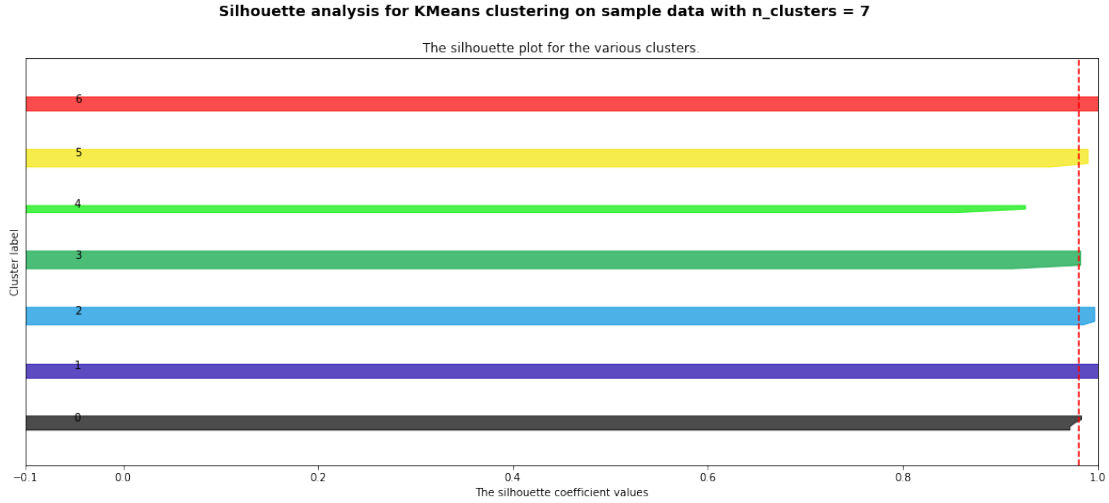


Fig. 4.4 Puntuació del mateix model de la figura anterior però seguin el model de Silhouette

Com es pot veure a partir de les figures anteriors, el fet de disposar de dos escales de puntuació ens va facilitar l'elecció del nombre òptim de clústers en casos en els quals si només miréssim un únic model seria molt més complicada.

Posteriorment, es va repetir aquest procés de clusterització agrupant segons el nombre d'activacions per hora del dia i també en base a la durada mitjana dels esdeveniments, tan a

nivell de dia com d'hora.

Finalment, es van comparar els paràmetres que s'havien obtingut en cadascun dels diferents models de clusterització i es van buscar similituds entre ells per tal de discernir quin eren els clústers de dades comuns i que per tant havien de ser agrupats.

El resultat final fou la creació d'un total de 5 grups de paràmetres:

- HC_SRC
- HC_T
- HSC_HEATING_DHW
- DRINK_WTR_T
- HSC_DHW_STG_T

A més, cadascun d'aquests clústers estava format per un total de 6 paràmetres:

- DELTA_TEMP
- ENERGY
- POWER
- RETURN_TEMP
- SUPPLY_TEMP
- VOL

A partir d'aquest punt, es va passar a treballar amb aquests grups de paràmetres a l'hora d'intentar resoldre els problemes plantejats en les fases prèvies del projecte.

### 4.5.5 Càlcul del COP amb permutacions

Ara que disposàvem de les dades més rellevants, es va decidir repetir l'estudi del COP de la bomba de calor. Aquest cop, es van agafar totes les fonts de consum tèrmic que pertanyien a un dels grups i es va fer la divisió entre totes les fonts de consum elèctric. Aquest cop sí, utilitzant les transformacions adequades per garantir que les unitats amb les que es treballava

eren les mateixes. D'aquesta manera, es van obtenir totes les permutacions per intentar garantir que s'obtingués el COP real de la bomba.

A continuació, es van analitzar els resultats de les divisions i es van eliminar aquelles resultats que no tinguessin sentit. Després d'aquest procés, van quedar un total de 4 possibles COPS.

Malauradament, al representar-los gràficament es va observar que encara que els resultats obtinguts es trobaven dintre d'un rang acceptable (COP d'entre 2 i 4 majoritàriament), existia una altíssima variabilitat en totes les lectures normals, fent que el COP passés de 1 a 7, de 7 a 3, etc. A més, en moments puntuals es disparava fins arribar a tenir valors de 100.

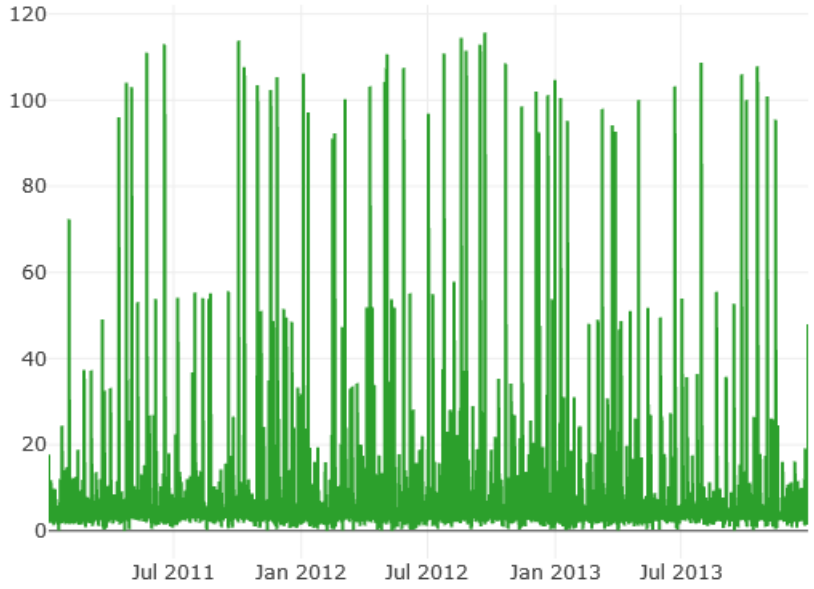


Fig. 4.5 COP corresponent a la divisió entre HSC_DHW_STG_T_POWER_ON i COMPRESSOR_E_POWER_ON

Evidentment, això no representa el funcionament normal de la bomba de calor, ja que apart de que s'hauria de mantenir relativament constant i deixant de banda els valors puntals de 100, els valors propers a 1 tampoc tenen sentit ja que una bomba de calor estàndard tendeix a tenir un COP mínim de 2 i rarament arriben a tenir valors de COP més grans de 4. Després d'observar aquestes dades, es va decidir deixar de banda aquesta aproximació.

### 4.5.6 Estudi de la DELTA_T

Finalment, abans de passar a treballar amb algorismes de detecció d'anomalies, es va fer una última prova per intentar determinar l'eficiència de la bomba de calor. Aquest cop, es va fer utilitzant la DELTA_T. La idea sent, que si la diferència entre la temperatura d'entrada i la de sortida del circuit disminuïa amb el temps, significaria que l'eficiència de la bomba estaria baixant.

De la mateixa manera que ens va passar amb el COP en la primera prova que es va fer, la DELTA_T varia amb el temps depenent de l'estacionalitat, per tant no s'observa de primeres cap patró que pogués indicar que la DELTA_T disminueix amb el temps. Els graf corresponent a la DELTA_T de HC_SRC es mostra a continuació.

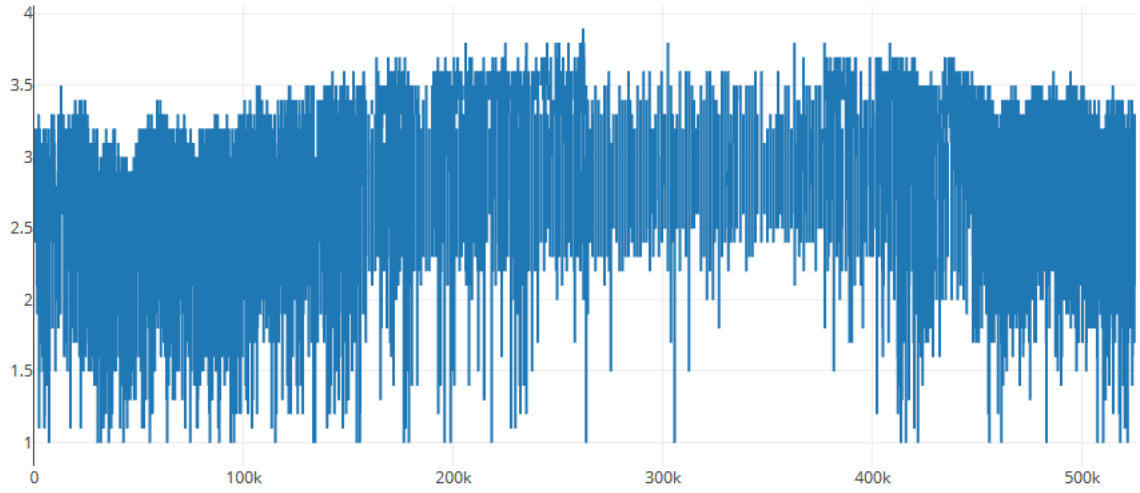


Fig. 4.6 Valor al llarg d'un any de HC_SRC_DELTA_T

## 4.6 Data Split

En aquesta secció, s'explica com es va determinar la millor manera de dividir les dades de les quals es disposava entre dades d'entrenament i dades de testeig de cara a passar-les al model que es va desenvolupar en l'apartat de Modelling.

Com que en el nostre projecte es tracta amb dades en *time series*, és a dir, dades ordenades en base a una data i una hora, els mètodes tradicionals de cross-validation no serveixen

ja que és necessari mantenir el ordre de les dades perquè sinó perdrien gran part del seu sentit.

És per això que es va decidir utilitzar un mètode no tan conegut de cross-validation que, ja que no te un nom oficial, s'ha decidit emprar el terme Rolling Cross-validation per referir-nos-hi, ja que és el terme que sembla tenir més acceptació. La idea darrere d'aquest mètode és relativament senzilla i el diagrama següent en dóna una visió molt clara:

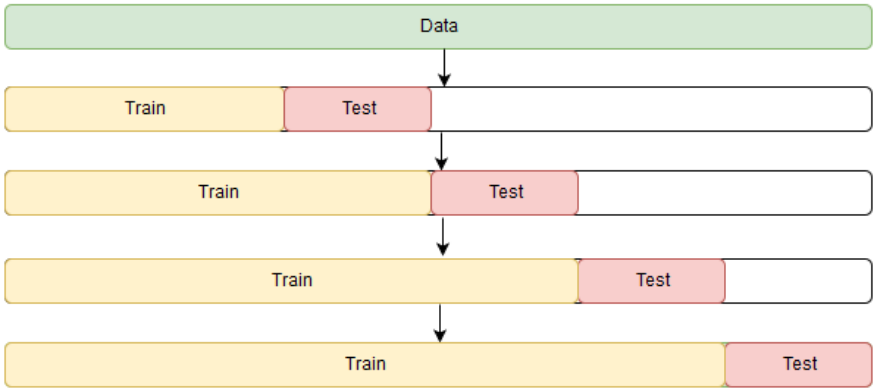


Fig. 4.7 Diagrama de funcionament de Rolling Cross-validation

Dividim el nostre data set de mida **s** en **n** particions iguals de mida **m**. A continuació entrenem el nostre model amb les dades que van de 0 fins a m i utilitzem les dades de m+1 a 2m com a test set per calcular l'error. Repetim el procés anterior però aquest cop entrenem amb les dades des de 0 fins a 2m i utilitzem les dades de 2m+1 a 3m de test. Repetim aquest procés de forma successiva fins que el grup d'entrenament estigui format per les dades de 0 a s-m i el grup de test per les dades de s-m+1 a s. Un cop finalitza el procés, es fa una mitjana dels diferents errors obtinguts i així obtenim la puntuació d'error final.

## 4.7 Modelling

En aquest apartat, és quan es crea un model i després l'entrenem, el validem i avaluem la seva eficiència. Per un motiu d'organització, en aquest apartat s'explicaran els algorismes que s'han utilitzat i en l'apartat de Deployment es discutiran els resultats obtinguts per cadascun.

En aquesta part del procés es van provar diferents algorismes abans de decidir per quin es crearia un model i es duria a terme tot el procés d'entrenament, validació i avaluació.

El primer que es va provar, va ser algorisme de detecció d'anomalies que utilitza Twitter, el qual està desenvolupat en R [4]. Malauradament, no es va aconseguir que funcionés correctament en el data set del projecte.

A continuació es va provar una opció molt naive: identificar els 0'5 i 99'5 percentils de les dades i marcar-les totes com a anòmales. Tot i que en alguns dels casos semblava funcionar relativament bé, quan el paràmetre era molt uniforme provocava que moltes dades normals fossin marcades com a anòmales.

A partir d'aquest punt, es va canviar de metodologia i es va començar a treballar amb algorismes de predicció més específics. La metodologia que es va utilitzar va ser la següent: generar una predicció, calcular el valor màxim i mínim que pot tenir el valor real i després, en cas de que el valor real es trobi fora del rang de valors que s'ha determinat, serà considerada anòmala.

Remarquem també que les proves que s'han realitzat s'han fet utilitzant el paràmetre HSC_HEATING_DHW_T_SUPPLY_TEMP. Això es deu a que aquest paràmetre és el que està connectat amb l'escalfador secundari de la bomba, el qual només s'activa quan una persona es dutxa. Com que aquest fet acostuma a succeir als vespre i es repeteix a diari, ens interessa buscar un algorisme que sigui capaç de detectar aquest esdeveniment i adaptar-s'hi. És per això que tots els algorismes seran executats amb aquest paràmetre.

El primer algorisme que es va provar amb aquesta nova metodologia fou l'algorisme d'exponential smoothing, el qual calcula el futur valor en base a una finestra dels últims x valors que ha tingut el paràmetre. L'algorisme utilitza un valor $\alpha$ per ponderar els valors que es trobin dintre de la finestra de forma exponencialment decreixent conforme la seva llunyania del valor actual.

El segon algorisme que es va provar fou l'algorisme de double exponential smoothing o algorisme de Holt. Aquest algorisme modifica l'algorisme d'exponential smoothing afegint un valor $\beta$ el qual pondera la component de pendent de la finestra i l'afegeix al càlcul del valor esperat.

El tercer algorisme fou l'algorisme de triple exponential smoothing o algorisme de Holt-Winters. De nou, aquest algorisme modifica el algorisme de Holt i afegeix un tercer

[4] Anomaly Detection with R: https://github.com/twitter/AnomalyDetection

component $\gamma$ que ens dóna una component d'estacionalitat. Per tal de determinar els paràmetres òptims de funcionament d'aquest algorisme es va dur a terme un proces de Rolling Cross Validation utilitzant l'error quadràtic mitjà com a funció de pèrdua.

Finalment, l'últim algorisme que es va provar fou un simple model de regressió lineal, en el qual el valors es calculaven utilitzant com a features les últimes 30 lectures del paràmetre. També es va entrenar el model utilitzant Rolling Cross Validation.

## 4.8 Deployment

Finalment, com que en el nostre projecte no es realitza cap desplegament del model que s'ha obtingut, en aquest últim apartat ens limitarem a presentar els resultats satisfactoris que hem obtingut de l'anàlisi de les dades.

### 4.8.1 Localització dels sensors en la bomba de calor

Mitjançant l'estudi de les dades i el seu comportament hem estat capaços d'identificar a quina part de la bomba pertany cadascun dels grup de paràmetres que hem obtingut anteriorment amb la clusterització.

Aquesta identificació ha estat possible gràcies a dos elements: el primer, va ser la col·laboració puntual d'un expert del domini el qual ens va indicar de forma general a quines parts de la bomba es podien correspondre cadascun dels paràmetres en base a la temperatura de l'aigua.

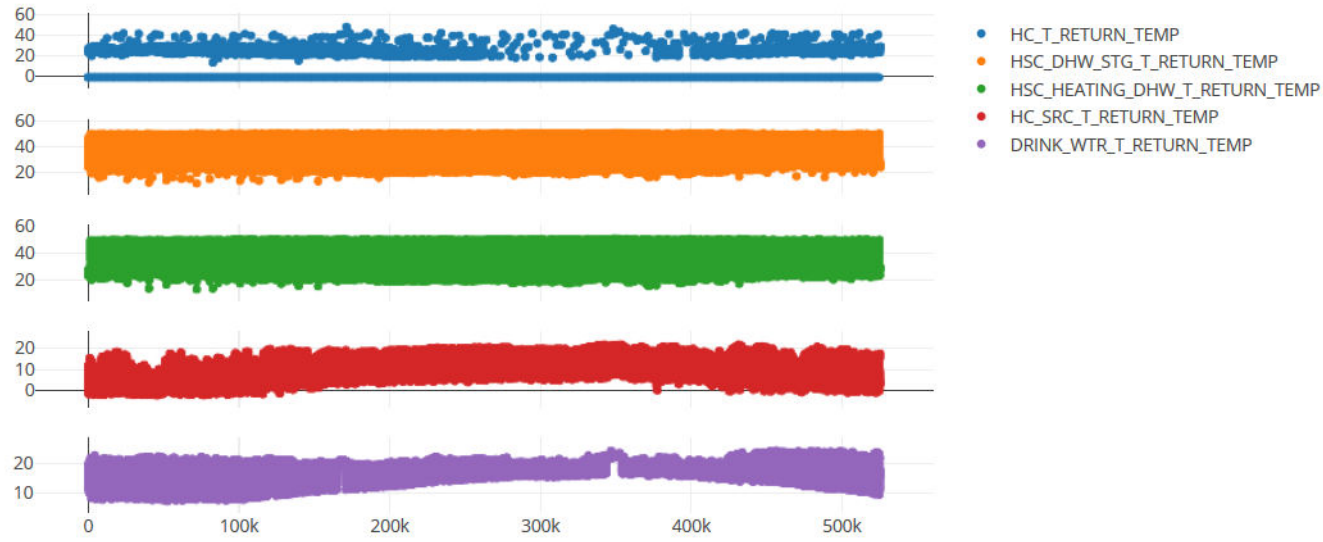


Fig. 4.8 Totes les dades de la temperatura de retorn per cada paràmetre

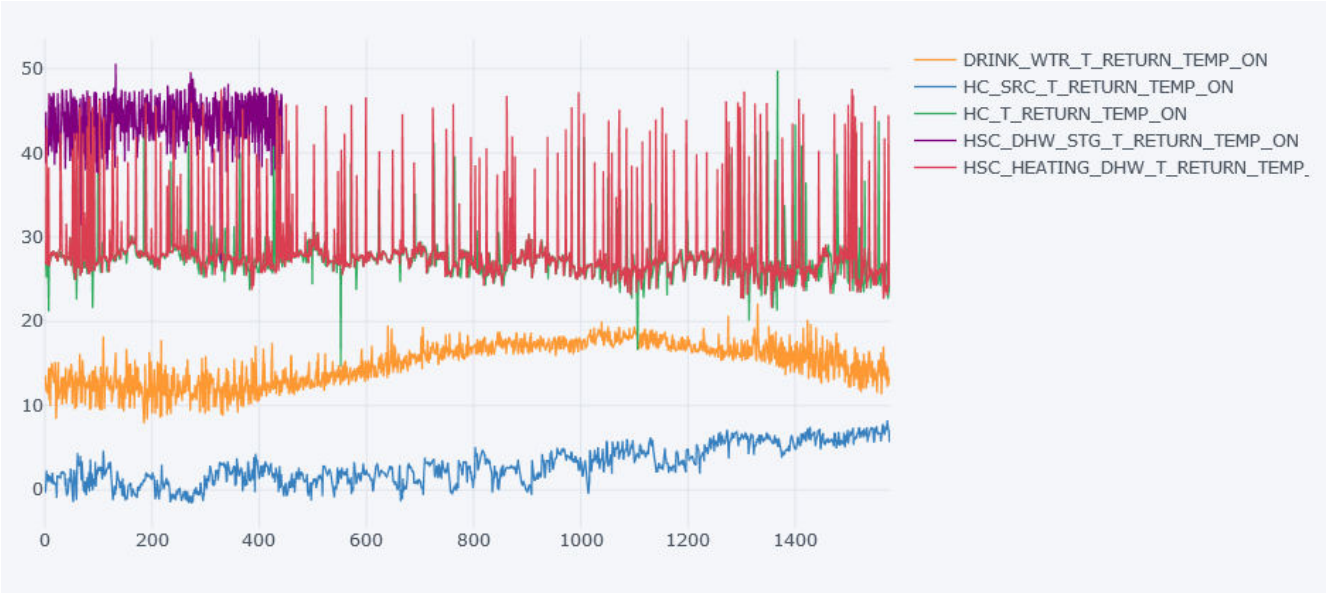


Fig. 4.9 Mitjana de la temperatura de retorn en els esdeveniments de cada paràmetre

Com es pot observar en la gràfica anterior, hi ha clara diferenciació entre les temperatures de funcionament de cadascun dels paràmetres i això permet a un expert determinar-ho fàcilment.

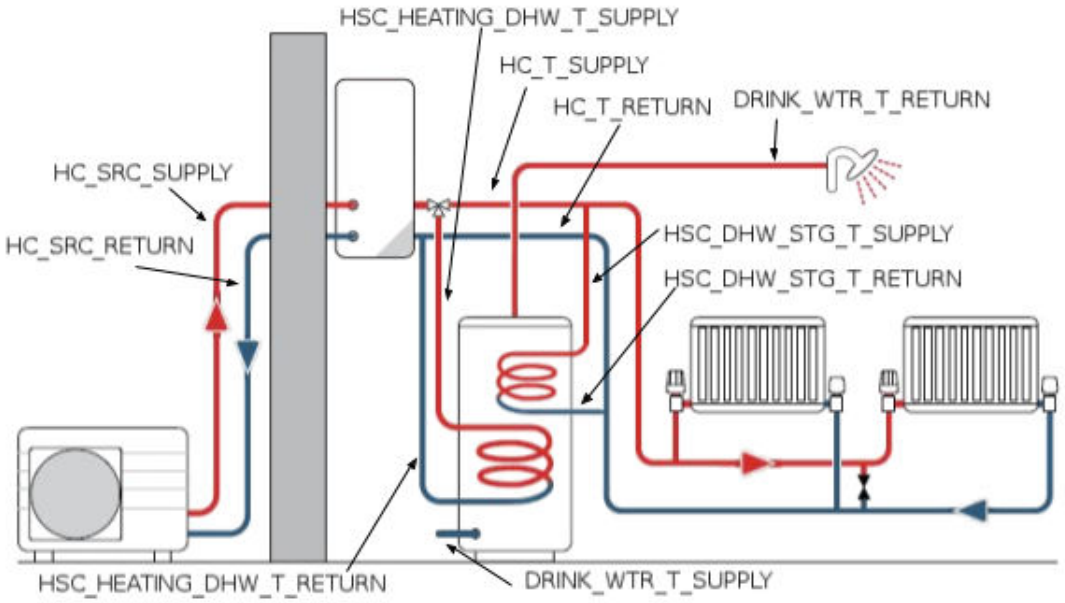


Fig. 4.10 Posició física dels paràmetres del data set en la bomba de calor

Tot i que l'expert només ens va orientar en quina podia ser la solució i no ens va arribar a donar cap confirmació dels nostres resultats, estem plenament convençuts de que la identificació que hem fet és correcta

### 4.8.2 Algorismes de predicció

Les representacions visuals d'aquests algorismes tenen totes un mateix estil, les línies discontínues vermelles indiquen els rangs inferior i superior dintre dels quals un valor és considerat normal. Els punts vermells, ens indiquen que un valor és una anomalia.

El resultat de l'algorisme d'exponential smoothing fou el següent:

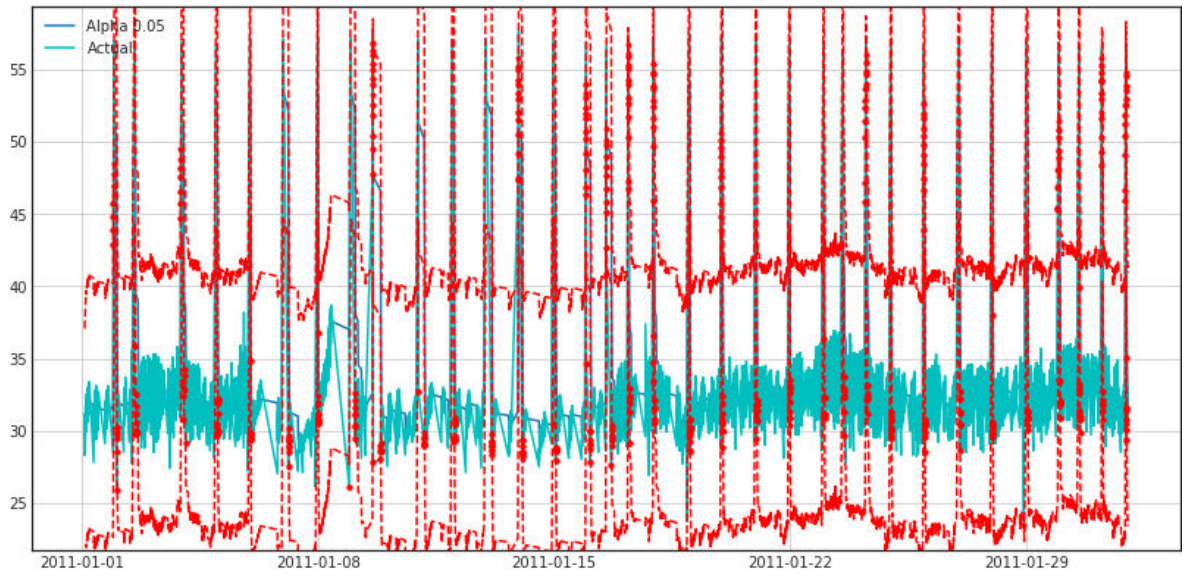


Fig. 4.11 Resultats obtinguts amb una $\alpha$ de 0.05

Com es pot observar, aquest algorisme no és capaç de tractar amb el pic diari i depenent del valor que donem a $\alpha$ també detecta com a anomalies comportaments normals del paràmetre.

El resultat de l'algorisme de Holt fou el següent:

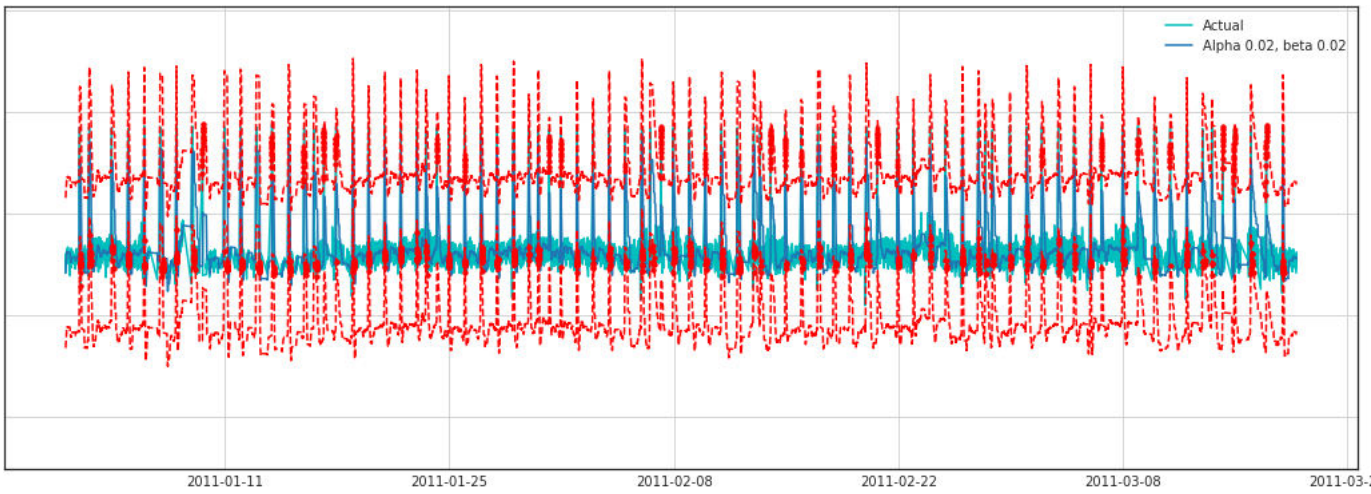


Fig. 4.12 Resultats de l'algorisme de Holt amb una $\alpha iuna\beta de 0.02$

Es pot observar una millora important respecte el single exponential smoothing pel que al tractament dels pics diaris, ja que quan ja porta una mica pujant l'algorisme s'adapta i ja no ho detecta com a anomalia. Per altra banda, han augmentat el nombre d'anomalies

detectades en la part central de les dades provocant que a nivell percentual, aquest segon algorisme detecti un major nombre d'anomalies que clarament no ho són.

El resultat de l'algorisme de Holt-Winter fou el següent:

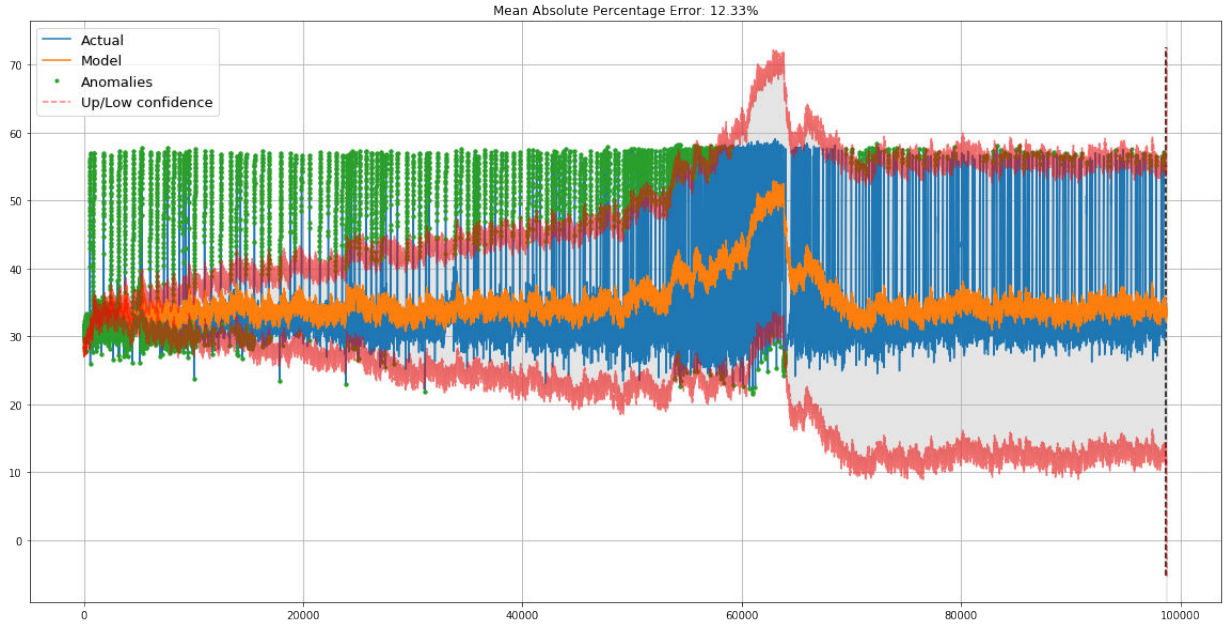


Fig. 4.13 Resultats obtinguts amb els valors òptims d'$\alpha, \beta i \gamma$

En aquest cas, sí que ha tingut lloc una millora substancial respecte els altres dos algorismes. Com es pot observar, el algorisme segueix sense detectar correctament el pics tot i utilitzar una component de periodicitat però sí que semblen seguir molt millor el funcionament normal del paràmetre.

Pel que fa al problema dels pics, després de fer certa recerca tot sembla indicar que l'algorisme de Holt-Winters només ofereix millores a través de la seva component estacional si aquesta és regular. En el nostre cas, el pic té lloc de forma regular als vespres, però una variació de mitja hora d'un dia a l'altre pot ser més que suficient per provocar que la component d'estacionalitat empitjori els resultats obtinguts enlloc de millorar-los.

I finalment, el resultat del model de regressió lineal ha estat el següent:

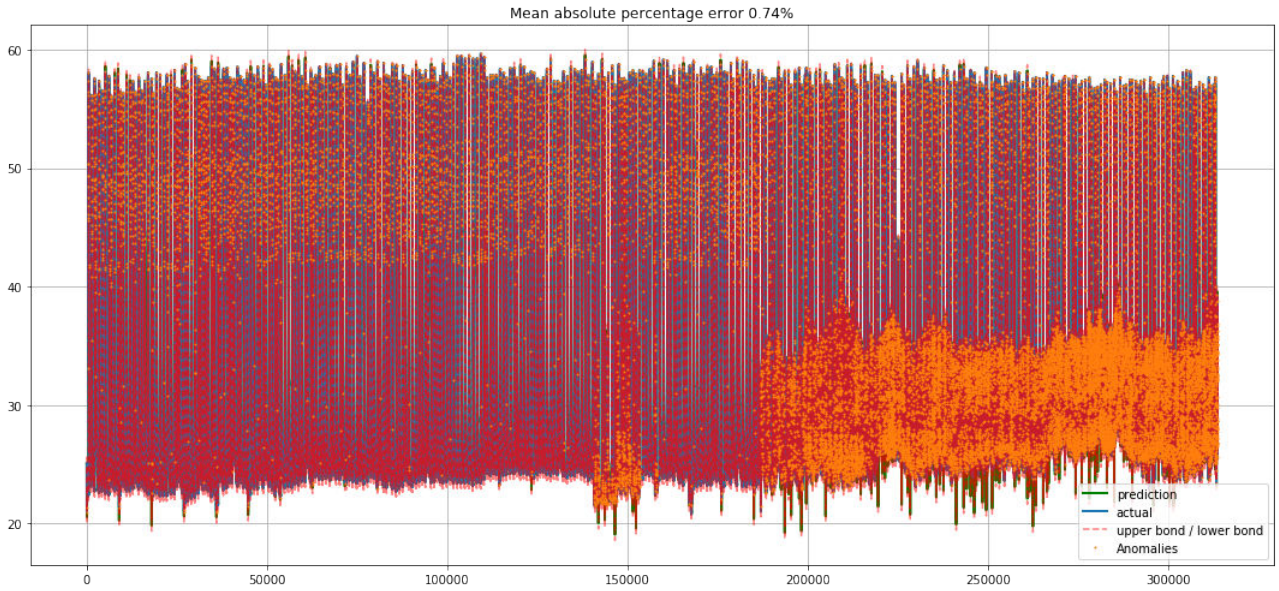


Fig. 4.14 Resultats obtinguts amb el model de regressió lineal

Es pot observar clarament que aquest model s'ha comportat molt millor que tota la resta, obtenint un percentatge d'error mitjà del 0.74 %, un error molt baix en comparació a la resta d'algorismes. No obstant, l'algorisme segueix tenint problemes, la qual cosa es fa palesa sobretot en la part final de la gràfica quan detecta un nombre d'anomalies massa elevat. No obstant, si s'afegeixen certs paràmetres addicionals per calcular la regressió lineal, crec que aquest model podria ser el que finalment ens permetés detectar anomalies.

# Capítol 5

# Conclusions i treball futur

Una vegada ja finalitzat aquest projecte, crec que a primera vista el resultats finals del projecte poden resultar enganyosos:

A nivell de resultats, no ha estat possible desenvolupar un sistema que ens permetés detectar senyals que ens indiquin que és necessari realitzar manteniment predictiu. Pel que fa als paràmetres d'eficiència de la bomba, no hem estat capaços d'identificar cap tipus d'empitjorament en el seu funcionament.

I pel que fa a la detecció d'anomalies, tot i que el model de regressió lineal que hem desenvolupat va pel bon camí i resulta molt prometedor i que considero que amb una mica més de treball seria capaç de detectar-les, el resultat final segueix sent que no ha estat capaç de detectar anomalies.

En aquest punt, m'agradaria fer un incís per tractar de forma més extensa un fet del qual s'ha parlat en diverses parts d'aquest projecte: la falta d'un expert del domini. Si ho analitzem des d'un punt de vista purament empresarial i d'eficiència, crec que de primeres és normal pensar que aquesta mancança probablement hagi comportat pèrdues de temps relativament importants, però aquesta afirmació perd valor un cop contextualitzada:

En aquest projecte s'ha treballat amb un dataset sobre el qual es disposava de molt poca informació més enllà de les dades. És per això que considero que la conclusió rellevant a la que s'ha d'arribar és que difícil extreure conclusions de dades amb poca contextualització. Potser sí que un expert del domini ho hauria facilitat, però no hauria deixat de ser un procés complex i llarg.

També és possible que els resultats haguessin estat molt més satisfactoris a nivell d'objectius complerts si un expert del domini hagués supervisat el projecte. Però també podria ser que simplement s'hagués arribat a les mateixes conclusions a les quals s'ha arribat, únicament hauria estat més ràpid.

A més, si ho analitzem des d'un punt de vista més educatiu, crec que resulta senzill veure que aquestes "pèrdues de temps" han resultat positives. I també estic convençut, de que aquest fet en ha portat a entendre molt millor què estàvem fent.

I es que, quin moment millor que aquest per experimentar amb llibertat de moviments, sense la pressió de tenir de complir objectius i així poder obtenir una experiència que molt probablement no es pugi repetir? Aquesta llibertat m'ha permès obtenir un bagatge elevat en l'ús tecnologies emprades i una experiència de la qual en podré treure partit en futurs projectes i que m'ajudarà a no repetir els errors que he comès en aquest projecte. Potser no s'han complert els objectius marcats, però sí que s'han assolit les dues principals premisses del projecte.

Primer de tot, hem estat capaços de dur a terme un projecte d'analítica de dades rigorós i el mes proper possible a un projecte del món real de principi a fi.

A més, crec hem creat una base de coneixement molt més extensa del que mai ho hauria estat si un expert hagués guiat les nostres accions. Hem après a actuar de forma autònoma i a adaptar-nos a la situació en la que ens trobàvem, i crec que aquest fet serà molt més valuós que qualsevol dels objectius que haguéssim pogut complir. Especialment, tenint en compte que el project LowUP comportarà treballar amb una gran diversitat de màquines que molt probablement requereixin desenvolupar o adaptar una solució per cadascuna d'elles.

És per tot això, que, tot i que els resultats no ho corroborin, considero que el projecte ha estat un èxit i els objectius més importants s'han complert.

Pel que fa a línies de treball futures, i seguint amb la figura de l'expert del domini, crec que seria factible identificar aquells paràmetres que realment determinen l'eficiència de la bomba de calor i el seu COP, la qual cosa ens donaria un altra forma de predir avaries i que podria ser combinada amb el model de regressió lineal per detectar anomalies més facilment.

Finalment, i tal com he comentat anteriorment, crec que el model de regressió lineal que hem desenvolupat el qual és molt prometedor, podria donar encara molts millors resultats amb certes modificacions, com podria ser afegir noves features, eliminar aquelles que no siguin rellevants o aplicant algun tipus de gradient boosting. De la mateixa manera, també crec que seria interessant explorar altres mètodes de predicció com podrien ser un model SARIMA o una Support Vector Machine o amb mètodes més complexes com podria ser una Xarxa Neuronal.

# Bibliografia


[1] Vasant Dhar. "Data science and prediction" in *Communications of the ACM*, 56(12), pp. 64–73, 2013.

[2] Eric Siegel. *Predictive Analytics: The Power to Predict Who Will Click, Buy, Lie, or Die*. Wiley, 2016.

[3] Anil Maheshwari. *Data Analytics Made Accessible*. Amazon Digital Services LLC, 2018.

[4] G. Kesavaraj and S. Sukumaran. "A study on classification techniques in data mining" in *2013 Fourth International Conference on Computing, Communications and Networking Technologies (ICCCNT)*, pp. 1-7, 2013.

[5] CRISP-DM 1.0, `https://www.the-modeling-agency.com/crisp-dm.pdf`. Accedit 14/3/18.

[6] Minds and Machines Europe Hackathon 2017 - EHPA & Fraunhofer Dataset, `https://github.com/PredixDev/minds-machines-europe/tree/master/Electrification%20Challenge/Heatpump%20Timeseries%20Dataset`

[7] Hashemian, H. M., and Bean, W. C. "State-of-the-Art Predictive Maintenance Techniques" in *IEEE Transactions on Instrumentation and Measurement*, pp. 3480–3492. 2011.

[8] Goldstein M, Uchida S. "A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data" in *PLoS ONE*, 11(4): e0152173, 2016.

[9] Blum, Avrim and Kalai, Adam and Langford, John. "Beating the Hold-Out: Bounds for K-fold and Progressive Cross-Validation" in *Proceedings of the Annual ACM Conference on Computational Learning Theory*, pp. 203-208, 1999.

[10] Christoph Bergmeir, José M. Benítez. "On the use of cross-validation for time series predictor evaluation" in *Information Sciences*, vol. 191, pp. 192-213, 2012.

[11] Open Machine Learning Course. Topic 9. Part 1. Time series analysis in Python, `https://medium.com/open-machine-learning-course/open-machine-learning-course-topic` Accedit 28/8/18

[12] Alex Reinhart. *Statistics Done Wrong: The Woefully Complete Guide*. No Starch Press, 2015.

[13] Wes McKinney. *Python for Data Analysis: Data Wrangling with Pandas, NumPy, and IPython*. O'Reilly, 2017.

# Apendix A

# Significat de les sigles dels paràmetres del Dataset

| Sigla | Unitat | Definició |
|---|---|---|
| ENERGY | kWh | Energia |
| E_POWER | W | Potència Elèctrica |
| T_POWER | kW | Potència Tèrmica |
| SUPPLY_TEMP | ºC | Temperatura de sortida |
| RETURN_TEMP | ºC | Temperatura d'entrada |
| DELTA_TEMP | K | Diferència de l'entrada a la sortida |
| AMB_TEMP (Ambient) | ºC | Temperatura Ambient |
| VOL (Volume) | l/h | Volum |
| OPT (Operation Time) | minut | Temps d'operació del cicle hidràulic anterior |
| TxE | ºC*kWh | Producte de la temperatura i l'energia |
| E | - | Elèctric |
| T | - | Tèrmic |
| ON | - | Durant el temps d'operació de la bomba |
| SPC_HEAT (Space) | - | Calefactor / Carregar el buffer d'emmagatzematge |
| STG (Storage) | - | Emmagatzematge |
| HSC (Heat Sink Circuit) | - | Circuit dissipador tèrmic |
| HP | - | Bomba de Calor |
| LP (Loading Pump) | - | Bomba de càrrega |
| HC (Heat Circuit) | - | Circuit d'escalfament |
| SPCH (Space heating) | - | Calefactor |
| DHW (Domestic Hot Water) | - | Aigua d'ús domèstic |

| Sigla | Unitat | Definició |
|---|---|---|
| EBH (Electric Backup Heater) | - | Caldera elèctrica de suport |
| SRC (Source) | - | Orígen/Font |
| FAN | - | Ventilador |
| BRINE_PUMP | - | Bomba de salmorra |
| WATER_PUMP | - | Bomba d'aigua |
| BUFFER | - | Buffer |
| HEAT | - | Escalafament |
| CONTROLS(_WO_PUMPS) | - | Controls (sense bomba) |

# Apendix B

# Scatter Plot dels paràmetres ON